Fuzzy C-Means

By
Balaji K
Juby N Zacharias

What is Clustering?

Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data.

Example:

The balls of same color are clustered into a group as shown below :

Thus, we see clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.

Fuzzy C-Means Introduction

- Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters.
- This method was developed by Dunn in 1973 and improved by Bezdek in 1981 and it is frequently used in pattern recognition.

Fuzzy C-Means Algorithm

It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , \qquad 1 \le m < \infty$$

where *m* is any real number greater than 1, *uij* is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *cj* is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy C-Means Algorithm

- **Step 1.** Initialize U=[uij] matrix, U(0)

- **Step 2.** At k-step: calculate the centers vectors C(k)=[cj] with U(k)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

Fuzzy C-Means Algorithm

- **Step 3.** Update U(k) , U(k+1)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

- **Step 4.** If || U(k+1) - U(k)||< threshlod then STOP; otherwise return to step 2.
- **Remark:** the computation of the updated membership function is the condition for the minimization of the objective function.

Remarks for C-means

- With fuzzy *c*-means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belonging to the cluster.
- The degree of being in a certain cluster is related to the inverse of the distance to the cluster.

Fuzzy C-Means

- By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set.
- Performance depends on initial centroids. For a robust approach there are two ways which is described below.

  1) Using an algorithm to determine all of the centroids. (for example: arithmetic means of all data points)

  2) Run FCM several times each starting with different initial centroids.

Case study

FUZZY C-MEANS CLUSTERING ON MEDICAL DIAGNOSTIC SYSTEMS

Introduction to case

- In this work, unsupervised clustering methods were performed to cluster the patients into three clusters by using thyroid gland data obtained by Dr.Coomans.
- Here, Fuzzy C-Means (FCM) and Hard C-Means (HCM) algorithms are used as an unsupervised clustering method to cluster the patients.
- As a result of clustering algorithms, patients' statuses are classified normal, hyperthyroid function and hypothyroid function.

Hard C-Means

- Hard C-Means clustering is also known as K-Means.
- The k-means algorithm partitions a collection of N vector into c groups
- The aim of the algorithm is to find the cluster centers(centroids) for each group. The algorithm minimizes a dissimilarity function

Hard C-Means Algorithm

- ***Step 1*.** Initialize the centroids ci,i=1,..c. This is typically achieved by randomly selecting c points from among all of the data points.
- ***Step 2.*** Determine the membership matrix U by Equation

$$u_{ij} = \begin{cases} 1 & if \| x_j - c_i \|^2 \leq \| x_j - c_k \|^2, for each k \neq i, \\ 0 & otherwise \end{cases}$$

- **Step 3.** Compute the dissimilarity function by using Equation below. Stop if its improvement over previous iteration is below a threshold.

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \|^2 \right)$$

Hard C-Means Algorithm

- **Step 4.** Compute new centroids using by Equation below Go to step 2.

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k$$

- The performance of the algorithm depends on the initial positions of centroids. So the algorithm gives no guarantee for an optimum solution.
- Hard k-means algorithm executes a sharp classification, in which each object is either assigned to a class or not.

Fuzzy C-Means

- Fuzzy C-means Clustering(FCM), is also known as Fuzzy ISODATA.
- The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.
- FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function.

Results

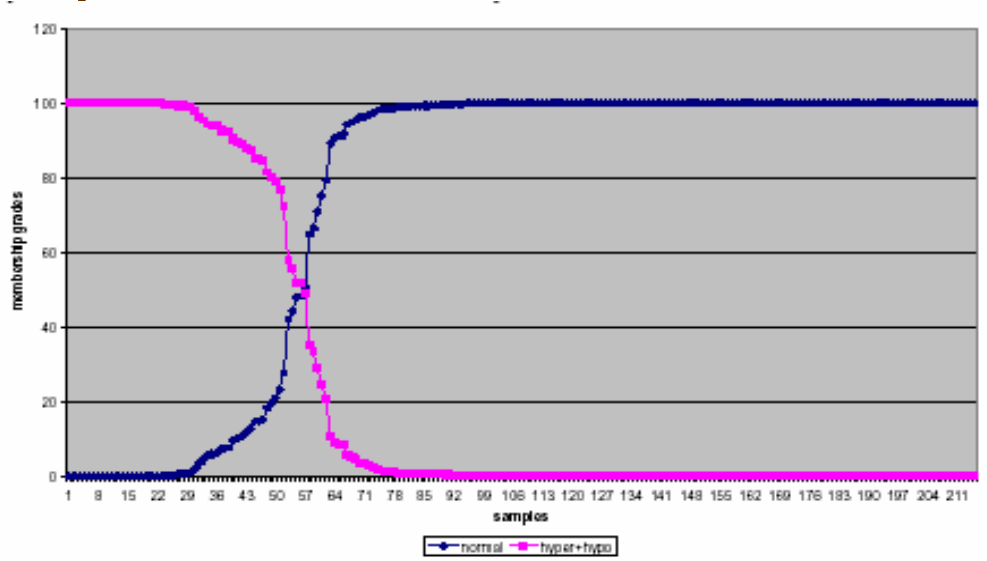
- Hard c-means algorithm is applied to thyroid gland data and 168 correct classified samples are obtained out of 215 samples. Fuzzy c-means give better results with 180 correct classified samples.

Table 1. Comparison of Classifiers

Classifier	# of Correct Classified Samples	Correct Classification Rate
Hard C-Means	168	78.1%
Fuzzy C-Means	180	83.7%

Results

- The application of Fuzzy C-Means causes the class membership to become a relative one and an object can belong to several classes at the same time but with different degrees. This is an important feature for medical diagnostic systems to increase the sensitivity.
- The graphical representation of the results is shown below



Results

- For the membership degrees close to 0.5 are the suspicious cases (shaded area in Table) to assign the sample to one cluster. Therefore fuzzy c-means clustering for medical diagnostic systems is more reliable than the hard one.

μ_3(hyper	μ_2(hypo)	μ_1(normal)	Samples
0	100	0	1
.....	.....	.....	2, 3....
80,3	0	19,6	51
0	79	21	52
76,7	0	23,3	53
72,2	0,1	27,8	54
55,6	0	44,4	55
51,9	0	48,1	56
0	49	50,9	57
33,7	0	66,3	58
29	0	71	59
.....	.....	.....	.....
0	0	100	215

Suspicious samples

Conclusion

- In medical diagnostic systems, fuzzy c-means algorithm gives the better results than hard-k-means algorithm.
- Fuzzy clustering methods can be important supportive tool for the medical experts in diagnostic.

References

1)FUZZY C-MEANS CLUSTERING ON MEDICAL DIAGNOSTIC SYSTEMS

2) www.endocrineweb.com/thyroid.html

3)Dr. Thangavelu's Seminar at IIITM-K

4)www.mathworks.com

5)http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletFCM.html


Thank U